

Georgia State University

ScholarWorks @ Georgia State University

World Languages and Cultures Invited Lectures
/ Talks

Department of World Languages and Cultures

2020

Invited lecture series on L2 pragmatics (2020): Lecture 5

Shuai Li

Georgia State University, sli12@gsu.edu

Follow this and additional works at: https://scholarworks.gsu.edu/wcl_ilt



Part of the [Modern Languages Commons](#)

Recommended Citation

Li, Shuai, "Invited lecture series on L2 pragmatics (2020): Lecture 5" (2020). *World Languages and Cultures Invited Lectures / Talks*. 5.

https://scholarworks.gsu.edu/wcl_ilt/5

This Lecture is brought to you for free and open access by the Department of World Languages and Cultures at ScholarWorks @ Georgia State University. It has been accepted for inclusion in World Languages and Cultures Invited Lectures / Talks by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.



Data collection methods & pragmatics assessment

October – December, 2020

- Session 1:
 - Data collection methods, pragmatics assessment, pragmatic competence.
- Session 2:
 - An empirical study on assessing speech acts in L2 Chinese.



Outline

Data collection and assessment

- Relationship between the two:
 - Assessment always involves data collection.
 - Data collection methods are not always suitable for purpose of assessment (e.g., practicality, reliability); e.g., field notes, recording of natural conversations, and think aloud protocols, etc.



Data collection methods

Pragmatics
assessment

	Elicited	Observational
Comprehension & Perception	Multiple choice Metapragmatic judgment Ranking Verbal protocol	Field notes Authentic discourse
Production	DCTs (oral, written) Role play Conversation task	Field notes Authentic discourse

Validity and validation

- Validity: the extent to which a test measures what it is intended to measure.
 - Construct validity: theoretical understanding of the construct to be measured.
 - The construct of pragmatic competence is multi-faceted, thus difficult to clearly define and operationalize.
- Two types of threat to construct validity in L2 pragmatics assessment.
 - Construct under-representation: e.g., to infer the ability of oral speech act ability based on request-making only.
 - Construct irrelevant variance: e.g., character recognition ability in oral speech act production.

Pragmatic competence – Recap (Lecture #1)

A pragmatics approach (a trait model).

- Pragmalinguistics & sociopragmatics / form-function-context mappings.
- Speech acts, routines, implicature, etc.

A discursive approach.

- Informed by the Conversation Analysis (CA) paradigm.
- Focus on interaction & co-construction of discourse.

A psycholinguistic approach (a trait model).

- Pragmatic knowledge & processing capacity.
- Performance appropriateness & performance fluency.

Assessments from the pragmatics perspective

- Hudson, Detmer, and Brown (1992, 1995).
 - First empirical effort to develop a test battery for assessing speech acts (i.e., requests, refusals, and apologies).
 - Developed speech act scenarios by drawing on Brown & Levinson's (1987) contextual variables: Power, social distance, and imposition.
 - Instruments for assessing speech acts: written DCT, oral DCT, multiple-choice DCT, role play.
 - Instruments for assessing confidence level in responding to DCTs and role plays.

Assessments from the pragmatics perspective

- Hudson, Detmer, and Brown (1992, 1995).
 - Written & oral DCT (sample item).
- You are a tourist in a large city. You have taken your film to a photo shop. When you go into the shop to pick up the pictures, the salesperson asks if you would like some coupons for more film developing. You do not need the coupons because you are leaving the city today.
You: _____.

Assessments from the pragmatics perspective

- Hudson, Detmer, and Brown (1992, 1995).
 - Multiple choice DCT (sample item).
- You work as a sales clerk in a department store. A customer is paying for an item and should get three dollars back in change. The customer asks that the three dollars be given in quarters, not dollar bills. You cannot give the change because you do not have enough quarters to spare. You say: _____.
 - A. I am sorry, but I don't have enough quarters.
 - B. No, I can't give you the quarters because I don't have enough.
 - C. I am sorry. I don't have enough quarters. I could give you two singles and four quarters.

Assessments from the pragmatics perspective

- Hudson, Detmer, and Brown (1992, 1995).
 - Role play (sample item from Cheng, 2011).
- Scenario: It's now around 4pm and you are leaving school. You want to stop by a bookstore and have heard that there is one named "Barns & Nobel" not far from school, but you do not know where it is. You are passing by the library and see your new classmate. You approach him/her to say some greetings. You two talk while walking together. The talk should include but may not be limited to the following points.
 - Start the conversation by greeting your friend and asking him/her about his/her classes. When it is most natural during the talk, compliment on his/her writing skills by referring to his/her essay published in the school news paper.
 - Ask for directions to get to the Barns & Nobel shop by bike / by car.
 - As what time the bookstore is closed today.
 -

Assessments from the pragmatics perspective

- Hudson, Detmer, and Brown (1992, 1995).
 - Self-assessment (sample item).
- You and a few of your co-workers are working on a special project. You are at a meeting in the office of the project leader. As you are reaching for your briefcase you accidentally knock over the project leader's umbrella which was leaning against the desk.

Rating: I think what I would say in this situation would be:

Very unsatisfactory 1---2---3---4---5 Completely appropriate.


Assessments from the pragmatics perspective

- Hudson, Detmer, and Brown (1992, 1995).
 - Holistic rating scale for scoring productions (written/oral DCTs, role plays).
 - Use of intended speech acts.
 - Conventionality of expression.
 - Amount of speech and information.
 - Formality, politeness, and directness.


These assessment dimensions reflect **pragmalinguistic** and **sociopragmatic** considerations.

Assessments from the pragmatics perspective

- Hudson, Detmer, and Brown (1992, 1995)'s test battery were later translated and adapted to assess Japanese (Yamashita, 1996) and Korean (Ahn, 2005) speech acts.
- Liu (2006, 2007) developed a test battery for assessing Chinese EFL learners' ability to make requests and apologies.
 - Notable methodological contribution in test item generation and selection, esp. scenario selection and distracter generation (for multiple-choice questions).



Assessments from the pragmatics perspective

- Roever (2005, 2006) pragmalinguistics assessment battery.
 - Focused on assessing pragmalinguistics of L2 English.
 - Expanded the construct of pragmatic competence: speech acts (DCT), implicature (multiple choice), and pragmatic routines (multiple choice).
 - The first web-based assessment tool.
- 

Assessments from the pragmatics perspective

- Roever's (2005, 2006) pragmalinguistics assessment battery.
 - Pragmalinguistic routines (sample item).
- Jane is at the beach and wants to know what time it is. She sees a man with a watch. What would Jane probably say?
 - A. Excuse me, can you say the time?
 - B. Excuse me, how late is it?
 - C. Excuse me, what is your watch show?
 - D. Excuse me, do you have the time? *

Assessments from the pragmatics perspective

- Roever's (2005, 2006) pragmalinguistics assessment battery.
 - Implicature (sample item).
- Jack is talking to his housemate Sarah about another house mate, Frank.

Jack: Do you know where Frank is, Sarah?

Sarah: Well, I heard music from his room earlier.

What does Sarah probably mean?

- A. Frank forgot to turn the music off.
- B. Frank's loud music bothers Sarah.
- C. Frank is probably in his room. *
- D. Sarah does not know where Frank is.

Assessments from the pragmatics perspective

- Roever, Fraser, and Elder (2014):
sociopragmatics assessment battery.
 - Focused on assessing sociopragmatics of L2 English.
 - A web-based assessment.



Assessments from the pragmatics perspective

- Roever, Fraser, and Elder (2014): sociopragmatics assessment battery.
 - Metapragmatic judgments of speech acts (request, apology, suggestion, and refusal); targeting **sociopragmatic awareness**, e.g.,
 - Jane needs to buy some stamps at the post office. She goes up to the counter, and the man behind the counter says: “How can I help you?”
Jane: “Hi, I am terribly sorry to bother you, but I was wondering if you might be so kind as to give me ten 50-cent stamps.”

Look at the final utterance. Do you think it is?
 - Very impolite / very harsh.
 - Not quite polite / soft enough.
 - Completely appropriate.
 - A little too polite / soft.
 - Far to polite / soft.

Assessments from the pragmatics perspective

- Roever, Fraser, and Elder (2014): sociopragmatics assessment battery.
 - Metapragmatic judgments of second pair parts (responses to offers, requests, compliments, etc.), with correction of inappropriate ones; targeting **sociopragmatic awareness and pragmalinguistic knowledge**, e.g.,

- Two friends are meeting at F1's new flat.

F1: So, do you like my new flat?

F2: It's very small. My flat is much bigger.

Is F2 responding appropriately? Yes / No

How should F2 respond? _____.

Assessments from the pragmatics perspective

- Roever, Fraser, and Elder (2014): sociopragmatics assessment battery.
 - Multiple-turn DCT targeting **pragmalinguistic and discursive abilities**, e.g.,
- Max and Julie are at a party. They don't know each other but happen to stand next to each other and start chatting.

Max: Hi, I'm Max.

Julie: _____.

Max: I'm good, thanks. Are you enjoying the party?

Julie: _____.

Max: Yeah, it's fun. So what line of work are you in, Julie?

Julie:

Assessments from the pragmatics perspective

- Roever, Fraser, and Elder (2014): sociopragmatics assessment battery.
 - Dialogue choice in terms of normativity of sequential organization, targeting **knowledge of discourse structure**, e.g.,
Which of the following is more successful in terms of communication skills used?

Dialogue 1:

Employee: Do you have a minute?

Boss: Sure.

Employee: I just wanted to ask you about this afternoon.

Boss: What's up?

Employee: Can I take one of the laptops and work from home?

Boss: Ok, how come?

Employee: The IT department is upgrading my computer today, but I need to finish these reports and email them out to everyone on the term by 4.

Boss: Sounds sensible. No problem. Let me know when you are leaving.

Employee: Thanks. Will do.

Assessments from the pragmatics perspective

- Roever, Fraser, and Elder (2014): sociopragmatics assessment battery.
 - Dialogue choice in terms of normativity of sequential organization, targeting **knowledge of discourse structure**, e.g.,

Dialogue 2:

Employee: Can I ask you a question?

Boss: Sure.

Employee: My child's school has been completely flooded after the storm this afternoon.

Boss: Oh. That sounds terrible.

Employee: The school is actually not that far from here.

Boss: Right.


Employee: It would be difficult to get someone else to pick him up.

Boss: You mean as a regular thing?

Employee: No, I need to leave early just for today.



Assessments from the pragmatics perspective

- Roever, Fraser, and Elder (2014): sociopragmatics assessment battery.
 - Targeted sociopragmatics in addition to pragmalinguistics.
 - The last two measures also tapped knowledge of discourse structure, thus leaning towards the discursive pragmatics perspective.
- 



Assessments from the discursive perspective

- Walters (2009): Conversation analysis (CA) informed pragmatics assessment.
 - Targeted pragmatic features informed by the CA literature:
 - Assessment responses: actions in which a speaker displays evaluations of event.
 - Compliment responses.
 - Pre-sequence responses: e.g., pre-invitation, see on next slide.

Assessments from the discursive perspective

- Walters (2009): Conversation analysis (CA) informed pragmatics assessment.
 - A sample item assessing comprehension of pre-sequence responses (e.g., pre-invitation):
Man: Hi Jane, this is Dick.
Women: Hi Dick.
Man: How ya doin-<uh what'r you guys doing? ←
Women: Well we're about to leave class, why?

In the conversation, what do you think the man will most probably do next?

- (a) Suggest going to class with the woman the others.
- (b) Offer to carry the woman's heavy book bag for her.
- (c) Explain that he had intended to make an invitation. *
- (d) Invite the woman and the other students to do something.

Assessments from the discursive perspective

- Walters (2009): Conversation analysis (CA) informed pragmatics assessment.
 - Walters' instrument showed relatively low reliability.
 - The instrument used CA-informed constructs, but his approach of assessing these constructs as isolated units does not fit the CA paradigm.
 - Walters (2007) assessed the production of similar CA-informed constructs in conversations and achieved better results.

Assessments from the discursive perspective

- Youn (2015): role play tasks (professor and classmate scenarios), esp. the rating criteria for assessing pragmatic production.
 - *Contents Delivery*: ability to deliver a turn appropriately and fluently.
 - *Language use*: Pragmalinguistics.
 - *Sensitivity to situation*: Sociopragmatics.
 - ***Engaging with interaction***: engagement in interaction (e.g., backchannel, clarification questions, acknowledgement tokens) and establishment of a shared understanding (i.e., the next turn is based on a good understanding of the previous turn).
 - ***Turn organization***: normative turn-taking conventions (e.g., preferred turns in a discourse such as “granting a request – thanking”).

Assessments from the psycholinguistic perspective

- Pragmatic knowledge and use/processing (Faerch & Kasper, 1984; Bialystok, 1993; Taguchi, 2012).
 - Pragmatic knowledge: Appropriateness and accuracy of pragmatic performance (i.e., pragmalinguistics and sociopragmatics); measures include, e.g.,
 - Ratings (for speech acts and routines).
 - Coding of strategies (for speech acts).
 - Accuracy scores (for implicature).
 - Use/processing: performance speed/fluency; measures include, e.g.,
 - Speech rate (for production).
 - Planning time (for production).
 - Response time (for comprehension).



Assessments from the psycholinguistic perspective

- The measures for assessing the use/processing component of pragmatic competence have not been integrated into pragmatics tests. They have mainly been used in SLA studies focusing on pragmatics.
 - Implicature comprehension (e.g., Taguchi, 2005, 2012; Taguchi, Li & Liu, 2013).
 - Speech act production (e.g., Taguchi, 2007; Li, 2014; Li & Taguchi 2014).



Additional data collection methods

- Additional methods used in L2 pragmatics research.
 - Metapragmatic judgement.
 - Ranking.
 - Conversation task.
 - Verbal protocols (think aloud).
 - Natural discourse / field notes.



Additional data collection methods

- Metapragmatic judgement: Shimamura (1993) for assessing pragmalinguistic and sociopragmatic awareness, e.g.,

Your friend from the mainland is visiting this week. You haven't seen her for a few years and this will be her first visit to Hawaii, so you have decided to take her around the island. But your car broke down and you do not want to spend a lot of money renting a car. Then you remember that your classmate who lives in the neighborhood just bought a new car last week. You decide to ask your classmate if you can borrow her car for the weekend. You say:

(1) Could I borrow your car this weekend if you are not using it? My car broke down. I will return it with a full tank of gas.

appropriate 1—2—3—4—5 not appropriate

(2) Could I borrow your car this weekend if you are not using it?

appropriate 1—2—3—4—5 not appropriate

(3) Is it your right to make the request to your classmate in this situation?

absolutely 1—2—3—4—5 not at all

Additional data collection methods

- Metapragmatic judgement: Schauer (2006) for assessing pragmatic vs grammatical awareness, e.g.,

<p>7. <i>Teacher Anna</i>, it's your turn to give your talk.</p> <p>! I can't do it today, but I will do it next week.</p>	<p><i>Was the last part appropriate/correct?</i></p> <p><input type="checkbox"/> <input type="checkbox"/> Yes No</p> <p><i>If there was a problem, how bad do you think it was?</i></p> <p>Not bad Very _____ : _____ : _____ : _____ : _____ : _____ : at all bad</p>
--	--

Additional data collection methods

- Ranking: Davis (2007) for assessing routine perception, e.g.,

Read the following responses and rank them in the order of preference.

1 = most likely

2 = might/possibly

3 = probably not

4 = never

You are working at a fast-food restaurant. You are helping a customer. The customer finishes ordering and wants to pay. You need to know if he wants to eat his meal in the restaurant or somewhere else. You say:

A) For here or to go?

B) To eat here or take away?

C) Where would you like to eat?

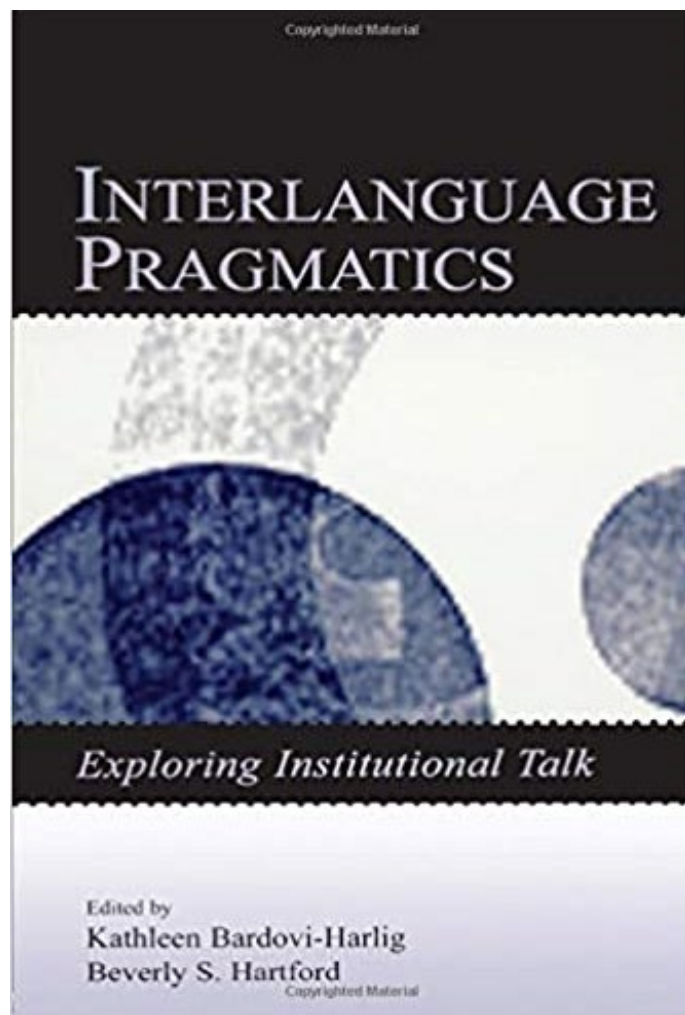
D) To eat inside or take outside?

Additional data collection methods

- Conversation task (Taguchi, 2015; Zhang, 2016).
 - Researchers put groups of participants together and provide topics for their joint conversation.
 - Different from role plays in that participants do not have to imagine a hypothetical role to play.
 - But this may restrict the type of contextual variables in designing assessment tasks (e.g., power, social distance, and imposition).
 - So far, very limited use in L2 pragmatics research so far, but good for assessing interaction.

Additional data collection methods

- Verbal protocols (think aloud).
 - To probe cognitive processes, strategies, and pragmatic awareness involved in task completion (Ren, 2014; Robinson, 1992; Xiao, 2017).
- Two types of think aloud techniques:
 - Concurrent: Participants verbalize their thinking processes as they complete a task.
 - Retrospective: Participants verbalize their thinking processes after they complete a task.
 - Timing is important: stimulated recall: Xiao (2017).
- Think-aloud requires participant training, individual sessions, and transcription of verbal reports before analysis (coding).



Additional data collection methods

- Natural discourse / field notes.
 - Issues regarding logistics, usefulness, standardization, and comparability.
 - Possibilities of collecting natural data.
 - Small scale, qualitative studies (e.g., Ishida, 2009).
 - Institutional discourse (Bardovi-Harlig, 2005): language classroom, tutoring sessions, advising sessions, and service encounters (e.g., Shivery, 2011).
 - Specific mode of communication, e.g., emails, phone voice messages, online chats (e.g., Belz & Vyatkina, 2003; Vyatkina, 2006).

Evaluating data collection methods

	Pros	Cons
Elicitation	<ul style="list-style-type: none">•Can fit specific research questions well.•Standardization & comparability.•(Relative) efficiency in data collection.	<ul style="list-style-type: none">•Validity issue: what is actually said vs. what people think they would say.
Observation	<ul style="list-style-type: none">•Authenticity & consequentiality.	<ul style="list-style-type: none">•Not easy to collect data for the purpose of a study.•Relatively time-consuming.•Reliability issue (accuracy in field notes, comparison across different situations, etc.).

Assessing L2 Chinese pragmatics

- 范香娟、刘建达 (2017) 外国留学生汉语中介语语用能力测量方法初探. 《语言教学与研究》第6期.
- Li, S. (2018). Developing a test of L2 Chinese pragmatic comprehension ability. *Language Testing in Asia*, 8, 1-23.
- Li, S., Taguchi, N., & Xiao, F. (2019). Variations in rating scale functioning in assessing speech act production in L2 Chinese. *Language Assessment Quarterly*, 16, 271-293.

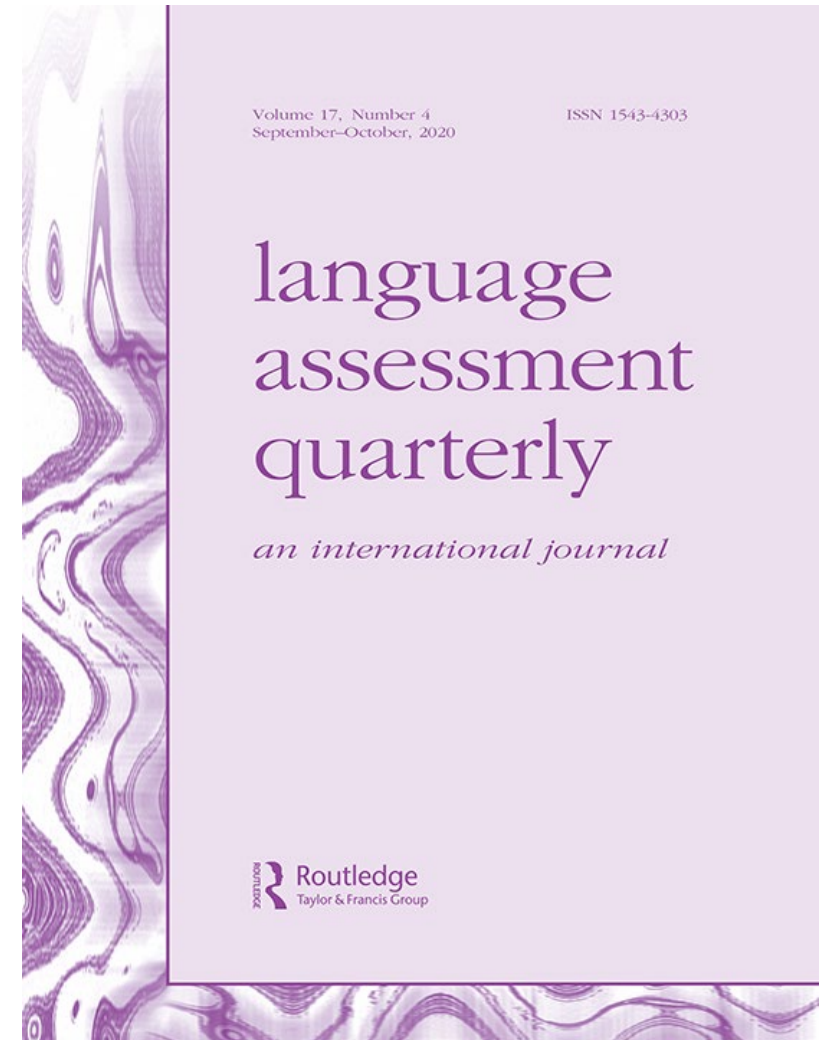
Let's take a
short break.

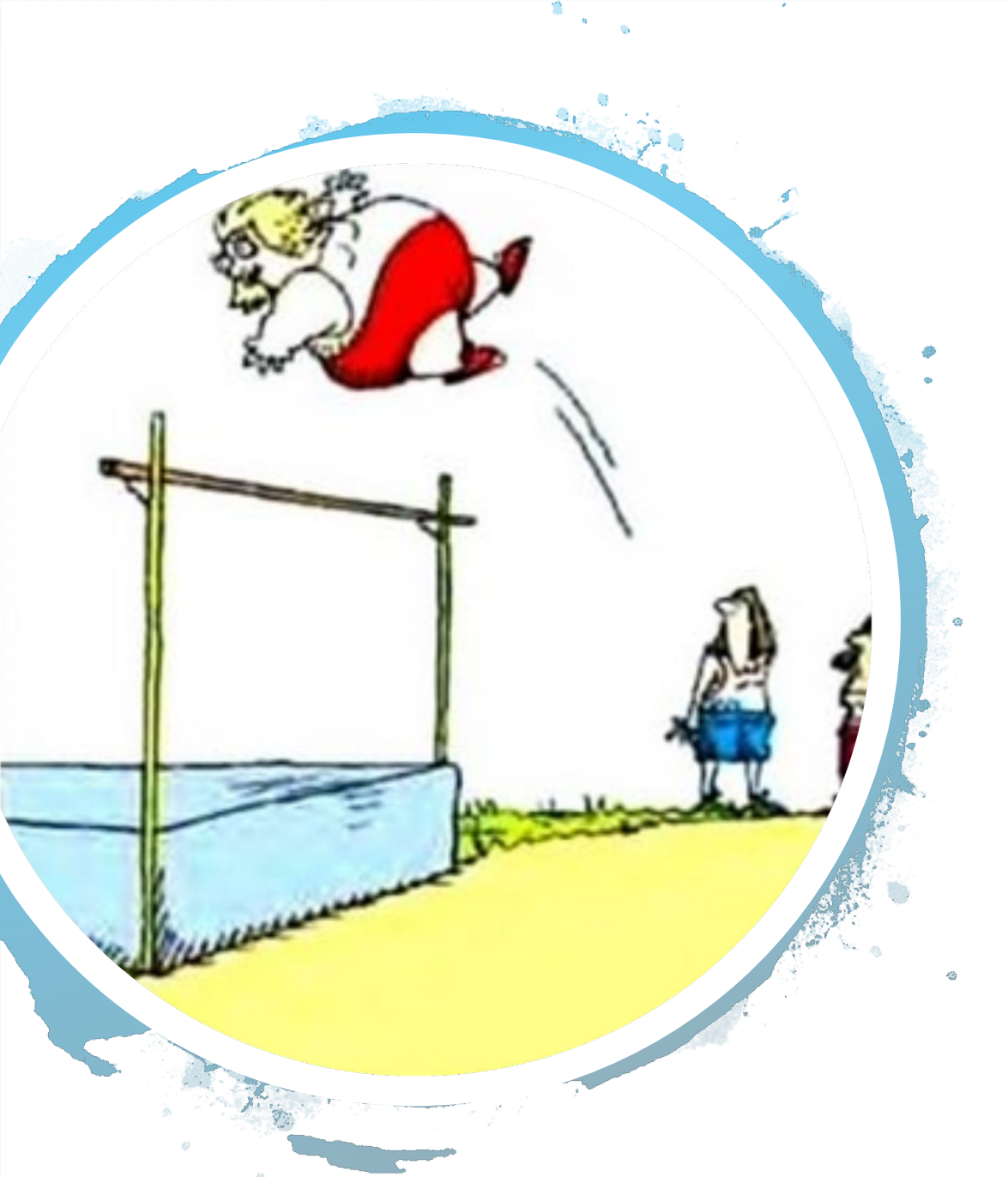




Session 2: An empirical study.

Li, S., Taguchi, N., & Xiao, F. (2019). Variations in rating scale functioning in assessing pragmatic performance in L2 Chinese. *Language Assessment Quarterly*, 16(3), 271–293.





Rasch model

- Rasch Model: a probabilistic psychometric model for measuring a latent construct through item responses.
- The difference between examinee ability and item difficulty determines the probability of the examinee succeeding in completing the test item.
- The difference is expressed in “logits”, in a logit scale. **The logit scale is an interval scale (a prerequisite for inferential statistical procedures).

Rasch model

- A hypothetical scenario:
 - A language ability test with 10 items, each worth 1 point.
 - Examinee A: 5/10 items correct → 5 points.
 - Examinee B: 5/10 items correct → 5 points.
 - Examinee A = Examinee B.
- However, what if A got the difficult items correct, and B only got the easy items correct?
 - Rasch Model takes item difficulty into consideration in estimating examinee ability.
 - So, Examinee A: 2.5 logits, and Examinee B: 1.9 logits.

Rasch model

- Rated responses:
 - Examinee ability.
 - Rater severity.
 - Item difficulty.

Measr	-Raters	+Examinees	-Items	S.1	S.2
2	+	+	+	(6)	(6)
		*			5

		*			
		***		5	

1	+	+	+	+	---
		**			

		*	RQ Air	---	

		*****	REF Presentation		

		****	RQ Exam	4	4
	Rater 2	*****	RQ Paper		

		*****	RQ Photo		
		*****	CR Cell phone		
*	0	*	*	*	*
		*****		---	
		****	REF New year		
		****	REF Tutoring		

	Rater 1	*****	CR Essay	3	---
		**	REF Dinner		

		*	CR Handwriting		
		*		---	3
-1	+	+	+	+	+
		*	CR Good Chinese		

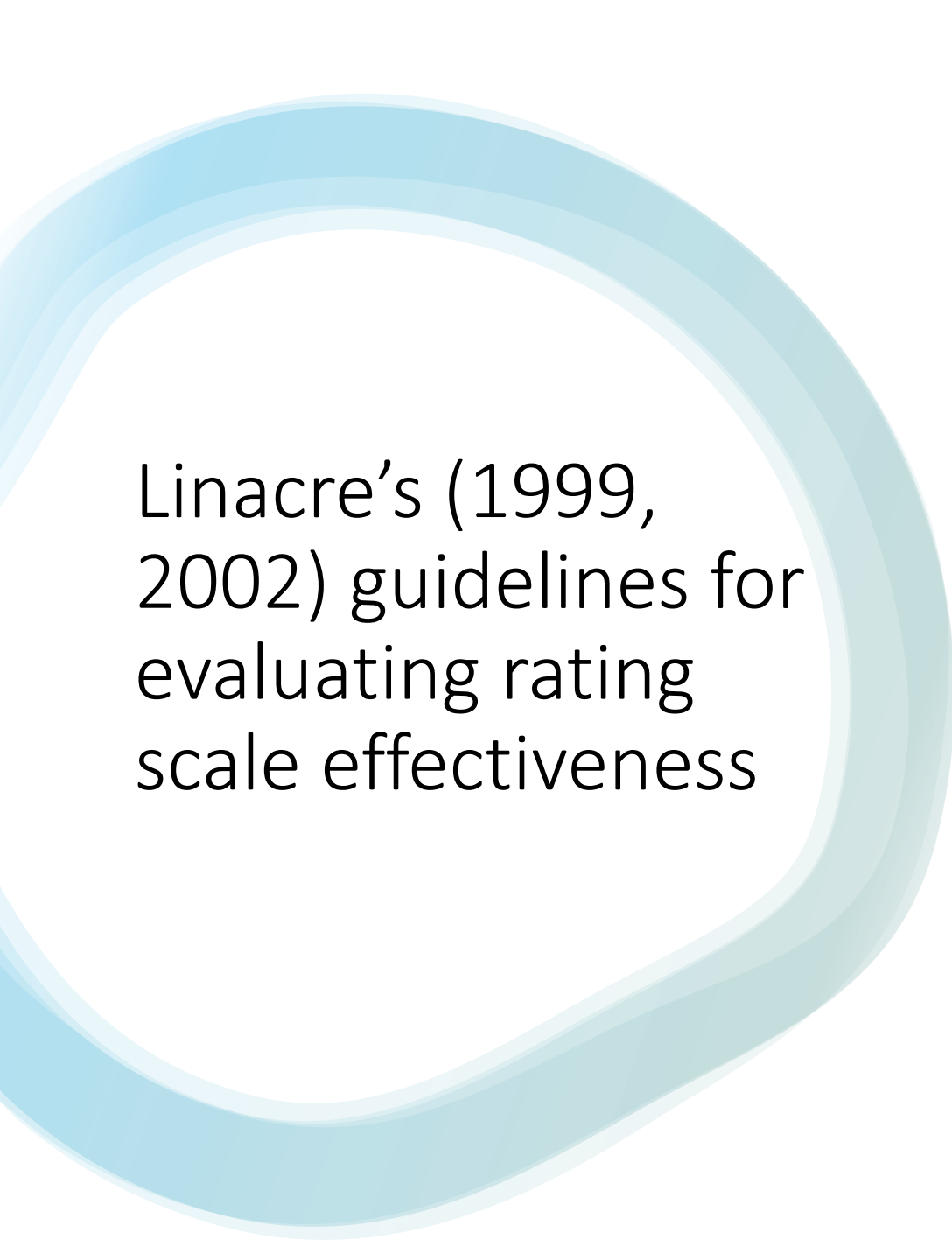
		*		2	---
		*			
		*			
-2	+	+	+	+	+
		*		---	2
-3	+	+	+	(1)	(1)
Measr	-Raters	* = 1	-Items	S.1	S.2

Background

- Major topics in L2 pragmatics assessment:
 - Instrument development and validation (e.g., Roever et al., 2014; Timpe-McLaughlin & Choi, 2017).
 - Rater behaviors (e.g., Liu & Xie, 2014; Taguchi, 2011).
 - Rating scale development and validation (e.g., Chen & Liu, 2016; Grabowski, 2013; Youn, 2015).
 - Differential item functioning (e.g., Roever, 2007).
 - Effects of examinee characteristics on test functioning (Youn & Brown, 2013; Roever, 2013).

Background

- Rating scales in L2 pragmatics research.
 - Widely used.
 - Self-developed or adapted.
 - Proper functioning rarely discussed.
- Initial evidence suggesting that rating scales may function differently according to raters, pragmatic features, and proficiency levels (e.g., Grabowski, 2013; Liu & Xie, 2014; Sydorenko, et al., 2014; Taguchi, 2011).
- This study adopted Linacre's (1999, 2002) guidelines to examine rating scale functioning in the context of assessing speech acts in L2 Chinese.



Linacre's (1999, 2002) guidelines for evaluating rating scale effectiveness

- **#1** Each rating scale category (hereafter category) should have minimally 10 observations.
- **#2** Regular distribution of category use frequency (e.g., even, unimodal, and bimodal distributions).
- **#3** Average (observed) measures increase monotonically with rating scale categories.
- **#4** Outfit mean-square (MnSq) statistic is below 2.0 for each category.
- **#5** Step thresholds (i.e., Rasch-Andrich thresholds) advance monotonically with rating scale categories.
- **#6** Observed measures of the rating scale categories approximate their expected values predicted by the Rasch Model.
- **#7 & #8** Increments of adjacent step thresholds fall into the 1.4-5 logits range. ** This study adopted the criterion of 1.0 logit as the minimum increment in step thresholds in order to be conservative.

Background

- Rationale of this study:
 - No study has investigated the effects of multiple contingent factors on scale functioning in one research design.
 - No study has explicitly adopted all of Linacre's (1999, 2002) guidelines as the basis for evaluating scale functioning.
 - The field has exclusively focused on English as the target L2.



Research Questions

- Is there any variation in rating scale functioning across:
 - (1) Raters?
 - (2) Speech acts?
 - (3) Proficiency levels?


Method

- Examinees.

- 109 American learners of L2 Chinese recruited from a study abroad program in Beijing, China (49 females, 60 males, age range: 19-23 years).
- Formal Chinese instruction: 1-7 years (Mean = 2.1 years).
- Proficiency: New HSK Level-4 , HSKK-Intermediate level.
 - Lower ($n = 54$): Mean = 186.27, range: 122.5 – 221.5, $SD = 25.24$.
 - Higher ($n = 55$): Mean = 270.44, range: 223.0 – 337.5, $SD = 32.32$.



Method

- Computerized Oral DCT.
 - To elicit speech act productions.
 - 12 items covering requests (k=4); refusals (k=4), compliment responses (k=4).
 - Raters.
 - 2 native Chinese speaker raters.
 - Trained in Chinese applied linguistics.
 - Familiar with L2 pragmatics research.
 - Over 10 years of Chinese language teaching experience.
- 

Method

- 6-point holistic rating scale simultaneously tapping:
 - Realization of communicative functions.
 - Situational appropriateness.
 - Grammaticality.



Scores	Descriptions
6 Excellent	<ul style="list-style-type: none"> ✧ Target communicative function fully realized ✧ Expression fully appropriate for a given scenario as judged by native speaker raters ✧ No or almost no syntactic/lexical errors
5 Very good	<ul style="list-style-type: none"> ✧ Target communicative function mostly realized ✧ Expression mostly appropriate for a given scenario as judged by native speaker raters AND/OR ✧ Limited syntactic/lexical errors (i.e., errors in peripheral lexical items, minor syntactic errors) that do not interfere with meaning and/or appropriateness
4 Good	<ul style="list-style-type: none"> ✧ Target communicative function somewhat realized ✧ Expression somewhat appropriate for a given scenario (e.g., verbosity, somewhat more direct and/or indirect than needed, use of uncommon semantic formula) as judged by native speaker raters AND/OR ✧ Syntactic and/or lexical errors tend to interfere with meaning and/or appropriateness
3 Fair	<ul style="list-style-type: none"> ✧ Target communicative function somewhat realized ✧ Expression clearly inappropriate (in terms of directness, formality, or semantic formula) for a given scenario as judged by native speaker raters AND/OR ✧ Notable syntactic and/or lexical errors (i.e., code switching, key lexical items) that clearly interfere with meaning and/or appropriateness
2 Poor	<ul style="list-style-type: none"> ✧ Target communicative function not realized ✧ Expression incomprehensible (due to serious phonological, syntactic/lexical error) OR ✧ Expression totally irrelevant to a given scenario (expression in this case may contain no, almost no, or some syntactic/lexical error) OR ✧ Expression is too limited for making judgment
1 Cannot evaluate	<ul style="list-style-type: none"> ✧ No response (opt out)

Method

- Procedures.

- Transcriptions were used for scoring.
- 2 raters jointly developed and revised the rating scale.
- Norming (3% of the data) followed by independent scoring.
- Out of the 1,308 possible responses (12 responses per examinee x 109 examinees), the final dataset included 1,303 responses for each rater, after excluding five invalid responses.
- 569 exact ratings (43.67%), 722 ratings (55.41%) differed by one point, and 12 ratings (0.92%) differed by either two or three points.
- Interrater reliability: Pearson's $r = .91$.

Method

- Data analyses:
 - Organized the dataset according to raters, speech acts, and proficiency levels.
 - Performed 3 separate three-facet Rasch Partial Credit Model analysis with *FACTES* 3.71.3 (Linacre, 2013).



RQ1 Results: Raters

- Satisfactory overall data-model fit.
 - Criteria: less than 5% of unexpected responses with absolute standardized residuals ≥ 2 , and less than 1% of unexpected responses with absolute standardized residuals ≥ 3 (Linacre, 2013, p. 162).
 - Findings: 118 (4.50%) residuals ≥ 2 , and 12 (0.46%) residuals ≥ 3 .

RQ1 Results: Raters

Measr	-Raters	+Examinees	-Items	S.1	S.2
2 +	+	+		(6)	(6)
		*			5

		*			
		***		5	

1 +	+	+			---
		**			

		*	RQ Air	---	

		*****	REF Presentation		

		****	RQ Exam	4	4
	Rater 2	*****	RQ Paper		

		*****	RQ Photo		
		*****	CR Cell phone		
* 0 *	*	*		* --- *	*

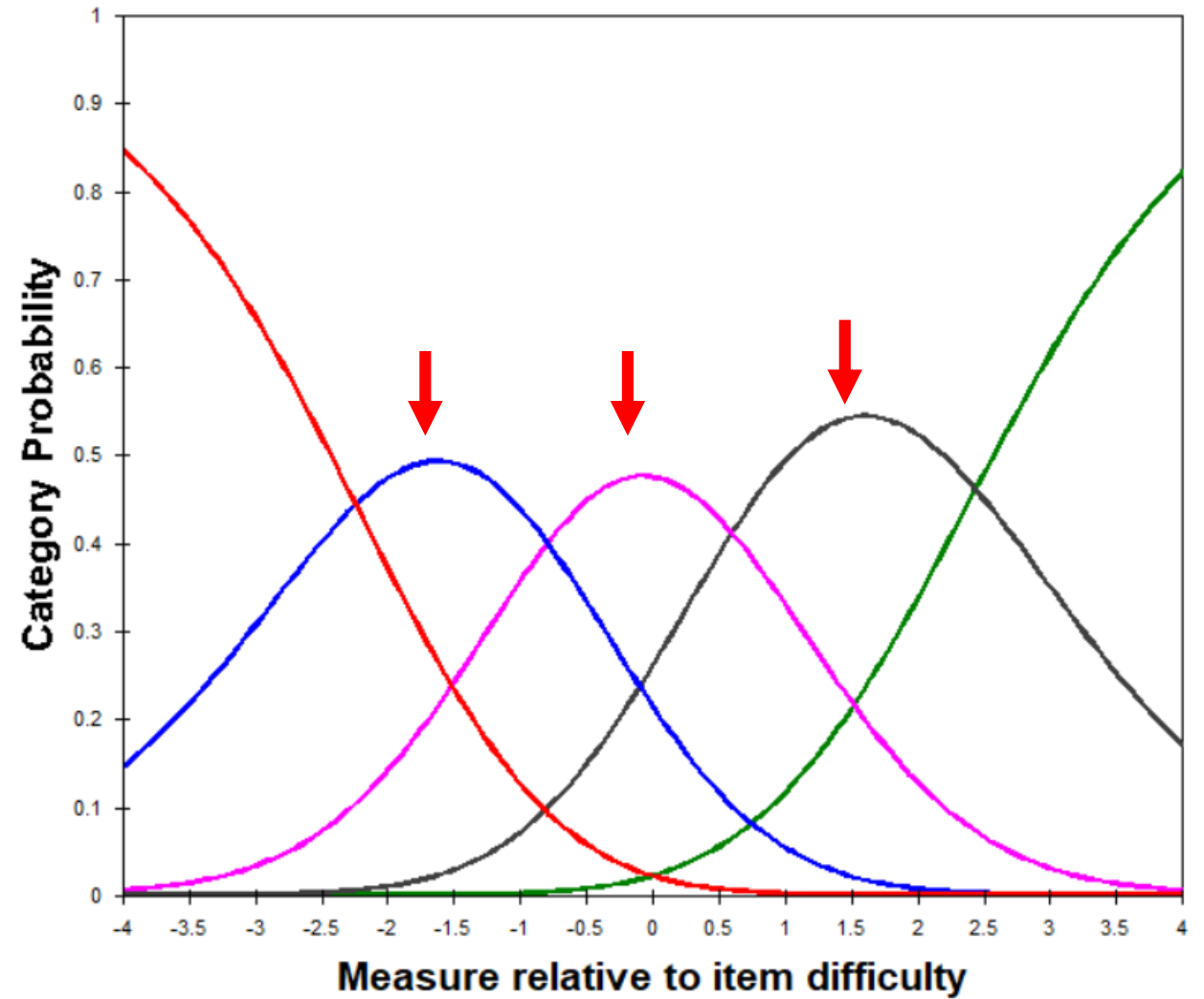
		****	REF New year	REF Tutoring	
	Rater 1	*****	CR Essay	REF Dinner	---
		**		3	

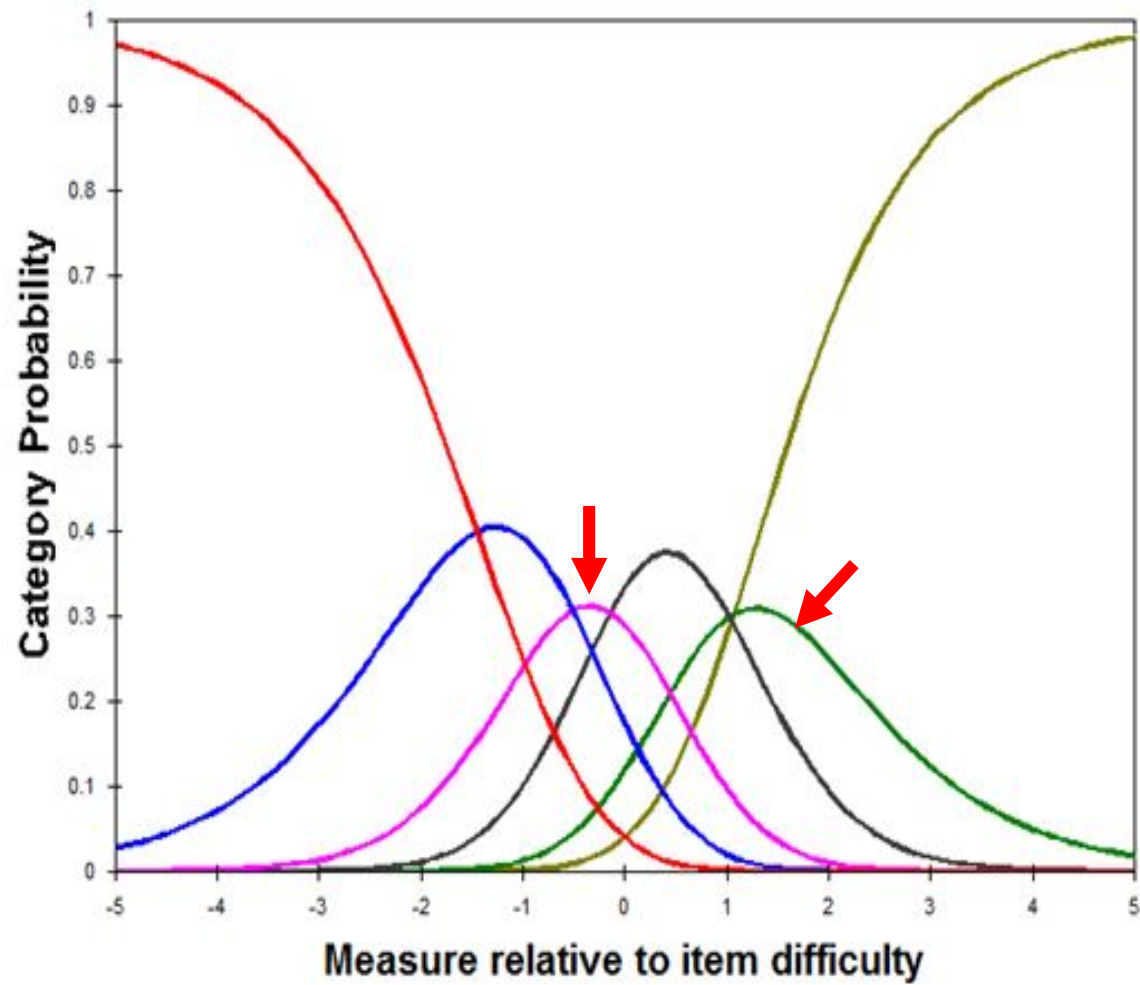
		*	CR Handwriting		
		*		---	3
-1 +	+	+			
		*	CR Good Chinese		

		*		2	---
		*			
		*			
-2 +	+	+			2
		*		---	
-3 +	+	+		(1)	(1)
Measr	-Raters	* = 1	-Items	S.1	S.2

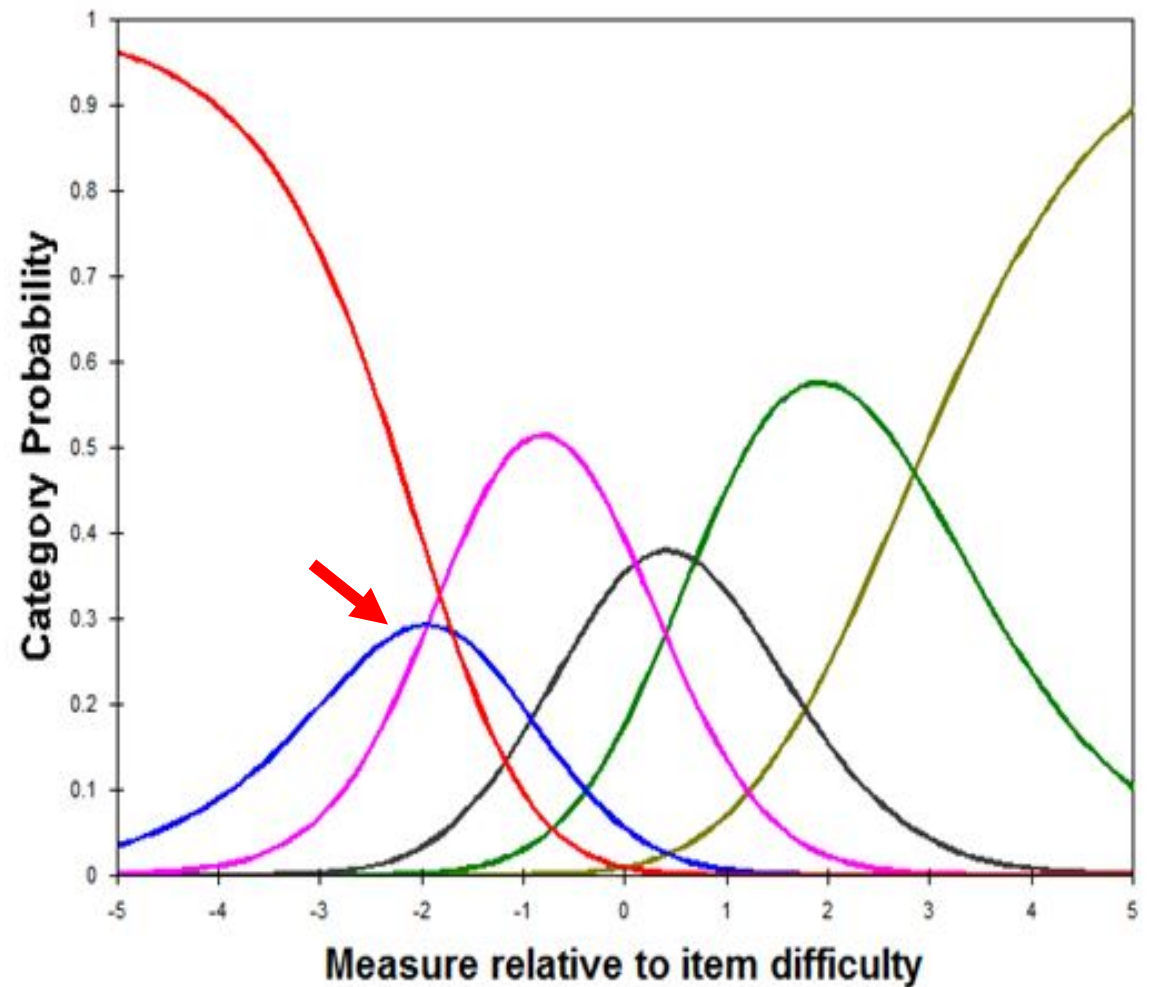
Guidelines	Rater 1	Rater 2
#1. Category frequency counts: ≥ 10	Yes, frequency count: 60 – 344	Yes, frequency count: 41 – 434
#2. Category distribution is regular	Yes, peaks at Category 4	Yes, peaks at Category 3
#3. Average measures Ascend monotonically	Yes	Yes
#4. Outfit MnSq < 2	0.9 – 1.1	0.8 – 1.2
#5. Step thresholds ascend monotonically	Yes	Disorder from Category 2 to Category 3
#6. Approximation between average measures and expected measures (in logits)	0.03 – 0.09	0 – 0.26
#7 & 8. Increment between step thresholds: range 1 – 5 logits	0.94 (Cat. 2 and 3) 0.65 (Cat. 3 and 4) 1.17 (Cat. 4 and 5) 0.04 (Cat. 5 and 6)	0.25 (Cat. 2 and 3) 2.05 (Cat. 3 and 4) 0.58 (Cat. 4 and 5) 2.16 (Cat. 5 and 6)

A more ideal
scenario
(from another
project of mine)





Rater 1



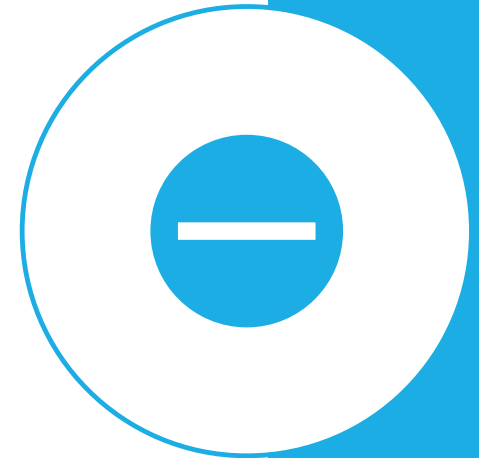
Rater 2

RQ 1 Discussion

- Previous studies (Taguchi, 2011; Walters, 2007) documented rater-induced variations in rating scale functioning due to raters' different linguistic, cultural, and professional backgrounds.
- But in this study, the 2 raters shared the same L1, came from the same country, attended identical educational programs, and worked in the same profession. They also developed the rating scale together and followed norming procedures typical in L2 pragmatics research. Yet, they still showed considerable variations in interpreting the same rating scale.
- The findings challenge the assumed homogeneity among L1 speaker raters with comparable backgrounds.
- Limitation: no rater thinking processes documented.

RQ2 Results: Speech acts

- Satisfactory overall data-model fit:
 - 100 (3.83%) residuals $\geq 2 \rightarrow$ below 5%
 - 13 (0.50%) residuals $\geq 3 \rightarrow$ below 1%



Measr	-Raters	+Examinees	-Items	S.1	S.2	S.3
2	+	+	+	(6) 5	(6) 5	(6) 5

		*				
		*****			5	
		**				5

		****		---		

1	+	+	+	+	---	+
		*****				---

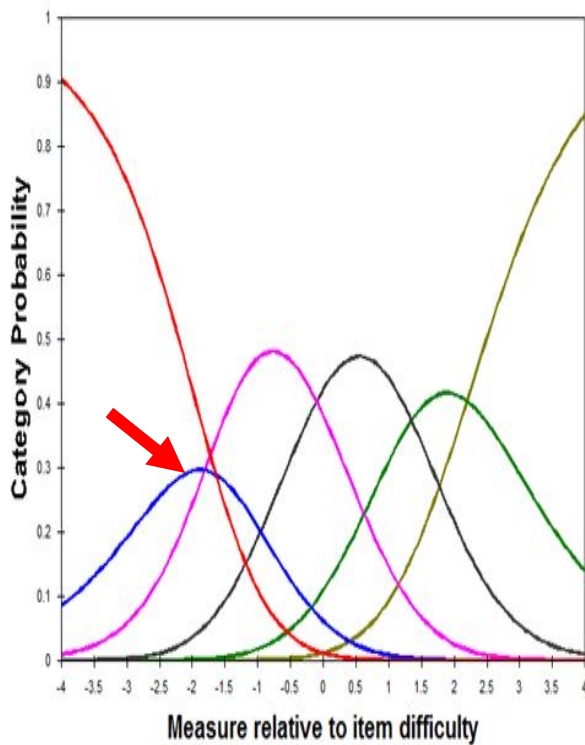
		*****	RQ Air			
		*****	REF Presentation	4		
		*****			4	4
		*****	RQ Exam			
		*****	RQ Paper			

	Rater 2	*****	RQ Photo			
		*****	CR Cell phone			---
* 0 *	*	***	* REF New year	*	---	*
		*****	REF Tutoring			
	Rater 1	**		---		
		**	CR Essay			3
		**	REF Dinner		3	
		*				
		*	CR Handwriting	3		---
-1	+	+	+	+	---	+
		***	CR Good Chinese			
				---		2
		*			2	
		*				
		*				
-2	+	+	+	+	+	+
		*			---	

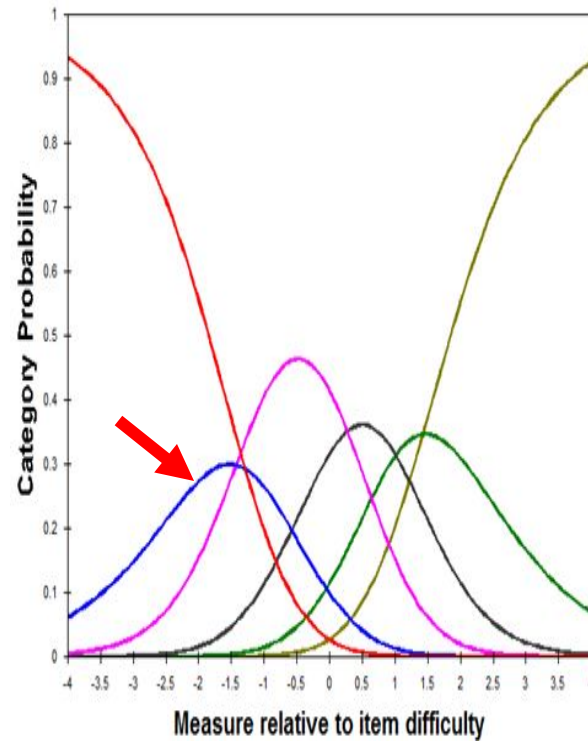
-3	+	+	+	(1)	(1)	(1)
Measr	-Raters	* = 1	-Items	S.1	S.2	S.3

RQ2 Results: Speech acts

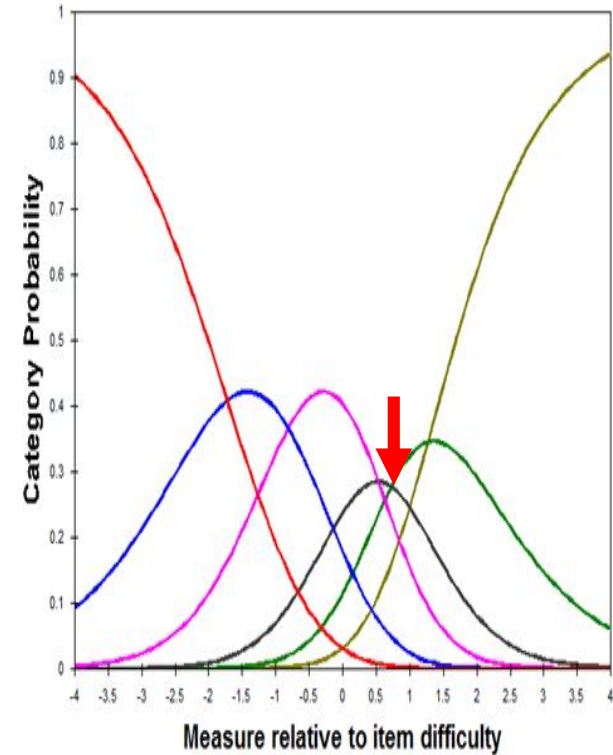
Guidelines	Compliment response	Refusal	Request
#1. Category frequency counts: ≥ 10	Yes, frequency count: 12 – 350	Yes, frequency count: 44 – 237	Yes, frequency count: 64 – 263
#2. Category distribution is regular	Yes, peaks at Category 4	Yes, peaks at Categories 3 & 4	Yes, peaks at Category 3
#3. Average measures Ascend monotonically	Yes	Yes	Yes
#4. Outfit MnSq < 2	1.0 – 1.5	0.8 – 1.4	0.6 – 1.0
#5. Step thresholds ascend monotonically	Disorder from Category 2 to Category 3	Disorder from Category 2 to Category 3	Yes
#6. Approximation between average measures and expected measures (in logits)	0.01- 0.31	0 – 0.20	0 – 0.10
#7 & 8. Increment between step thresholds: range 1 – 5 logits	0.17 (Cat. 2 and 3) 1.71 (Cat. 3 and 4) 1.45 (Cat. 4 and 5) 0.85 (Cat. 5 and 6)	0.12 (Cat. 2 and 3) 1.69 (Cat. 3 and 4) 0.75 (Cat. 4 and 5) 0.41 (Cat. 5 and 6)	0.91 (Cat. 2 and 3) 1.34 (Cat. 3 and 4) 0.19 (Cat. 4 and 5) 0.55 (Cat. 5 and 6)



Compliment responses



refusals



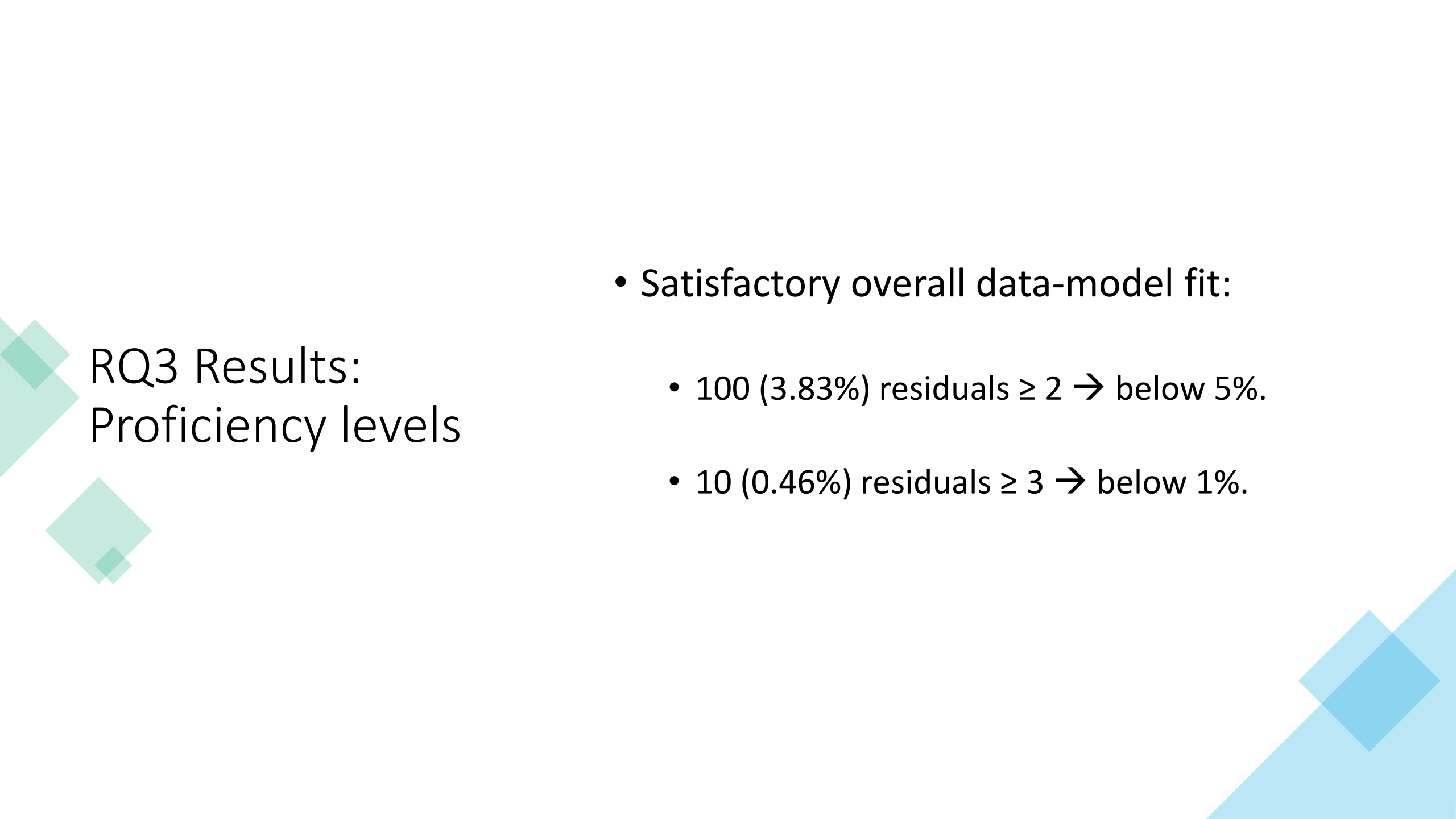
requests

RQ2 Discussion

- Varied difficulty of the three speech acts: Requests > refusals > compliment responses.
- The same generic rating scale categories may entail different substantive meanings according to the characteristics of specific targeted pragmatic features, causing variations in rating scale functioning. See examples on next slide.

RQ2 Discussion

- Scenario #1 (Compliment response):
 - *You wrote an essay about your travel experience and submitted to Professor Xiao's class. Today, you meet him in the hallway and you start to talk to each other. During your conversation, Professor Xiao says: "Oh, by the way, I read your essay and it is really interesting." What would you say to him? (Essay)*
 - Sample Response 1: “谢谢， 没关系。” (“Thanks, it doesn't matter.”) (a score of 3).
- Scenario #2 (Refusal):
 - *You come to Professor Sun's office to discuss a few questions with him. Before you leave, he invites you to a dinner party on New Year's Eve, but you cannot go. What would you say to Professor Sun? (New Year)*
 - Sample Response 2: “对不起， 我不去。” (“Sorry, I am not going.”) (a score of 3).



RQ3 Results: Proficiency levels

- Satisfactory overall data-model fit:
 - 100 (3.83%) residuals $\geq 2 \rightarrow$ below 5%.
 - 10 (0.46%) residuals $\geq 3 \rightarrow$ below 1%.

RQ3 Results: Proficiency levels

Measr	-Raters	+Examinees	-Items	S.1	S.2
2	+	+		(6)	(6)

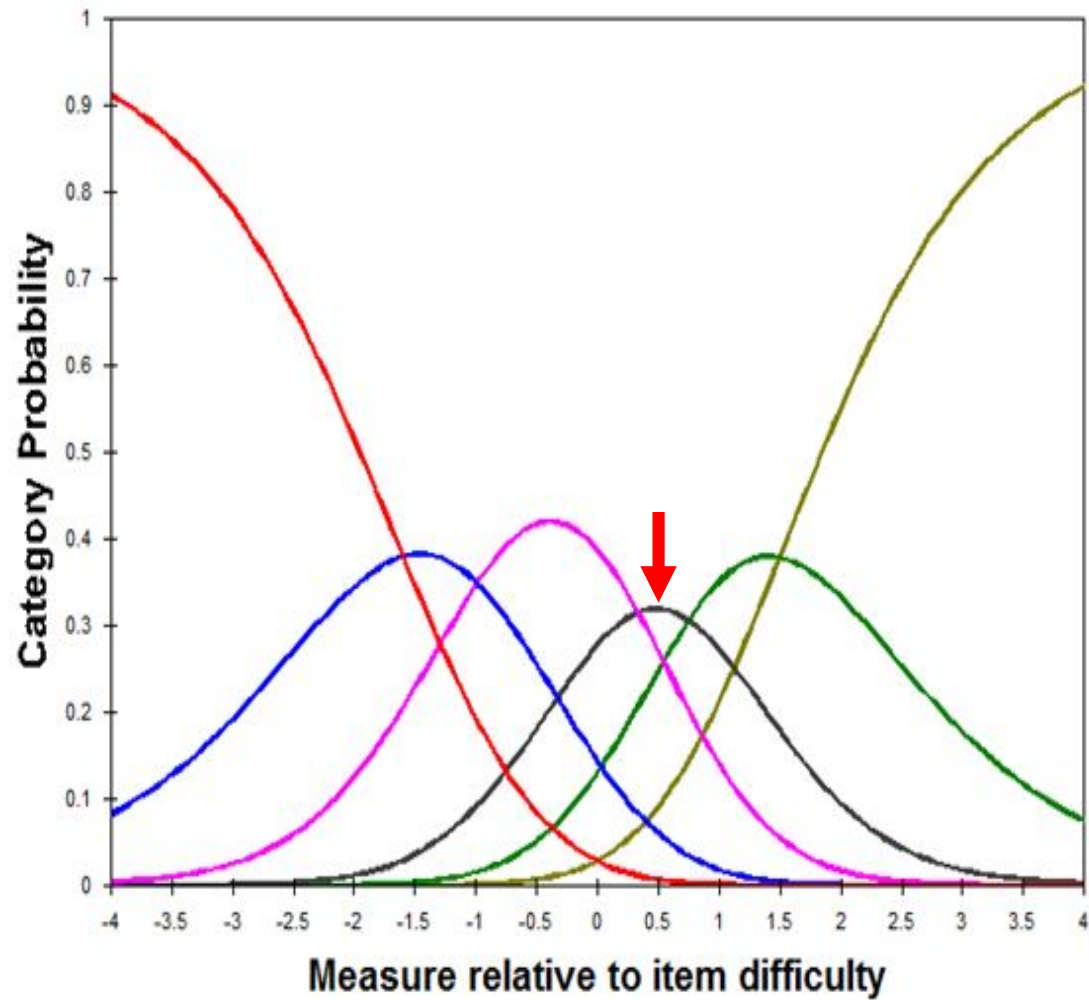
		**			5
		*			
		***		5	

		**			
1	+	+		+	---
		**		---	
		*****	RQ Air		
		*****	REF Presentation		
		*****	RQ Exam	4	4
		*****	RQ Photo		
	Rater 2	*****		---	

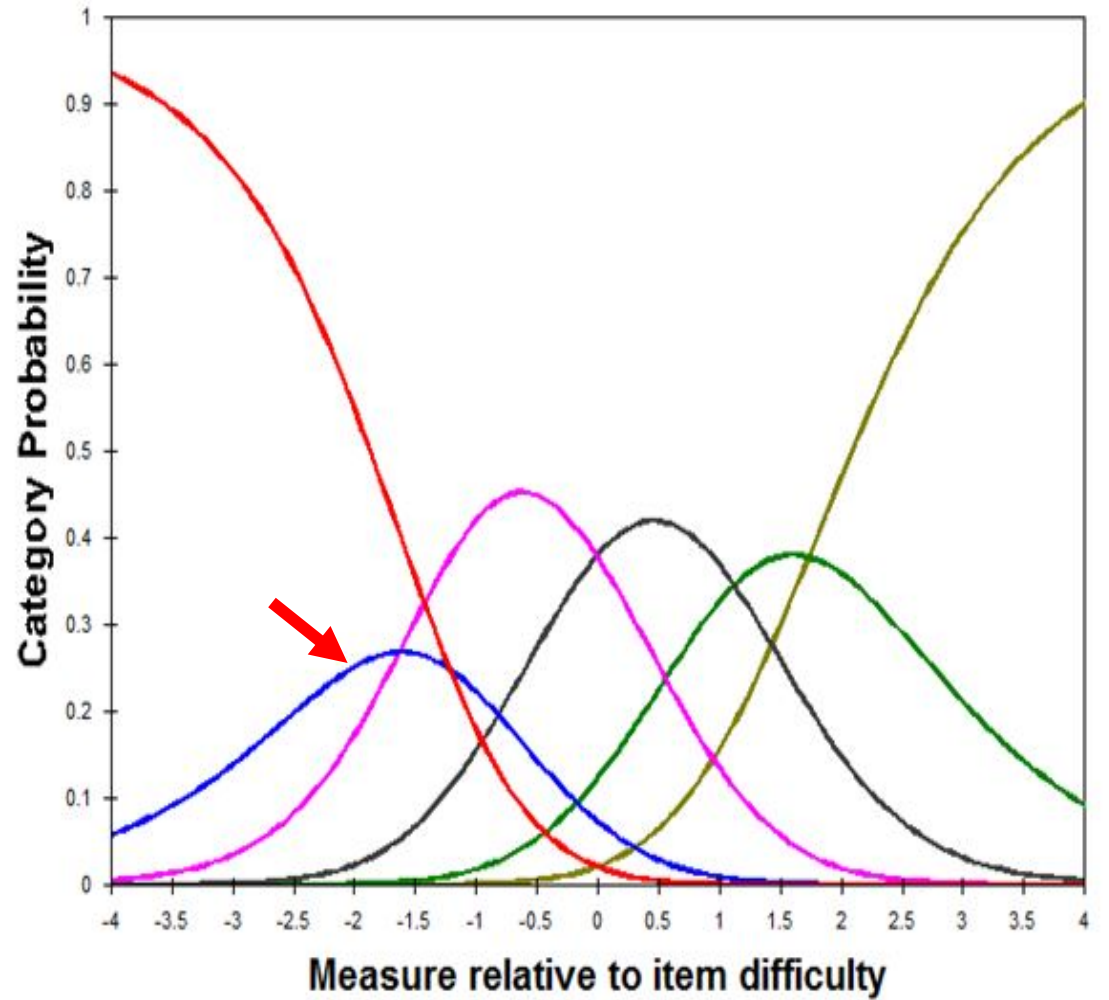
*	0	*		*	*
		*****	CR Cell phone		
		*****	REF New year		---
	Rater 1	*			
		****	CR Essay	3	

		*	CR Handwriting		3
		*			
		*			
		*	CR Good Chinese	---	
-1	+	+		+	---
		**			
		*			
		*		2	
		*			2
-2	+	+		(1)	(1)
Measr	-Raters	* = 1	-Items	S.1	S.2

Guidelines	Lower proficiency	Higher proficiency
#1. Category frequency counts: ≥ 10	Yes, frequency count: 96 – 360	Yes, frequency count: 24 – 429
#2. Category distribution is regular	Yes, peaks at Category 3	Yes, peaks at Category 4
#3. Average measures Ascend monotonically	Yes	Yes
#4. Outfit MnSq < 2	0.8 – 1.0	1.0 – 1.3
#5. Step thresholds ascend monotonically	Yes	Disorder from Category 2 to Category 3
#6. Approximation between average measures and expected measures (in logits)	0 – 0.2	0.01 – 0.3
#7 & 8. Increment between step thresholds: range 1 – 5 logits	0.61 (Cat. 2 and 3) 1.32 (Cat. 3 and 4) 0.42 (Cat. 4 and 5) 0.75 (Cat. 5 and 6)	0.41 (Cat. 2 and 3) 1.62 (Cat. 3 and 4) 1.13 (Cat. 4 and 5) 0.61 (Cat. 5 and 6)



Lower



Higher

Implications

- Proper functioning of rating scales in specific pragmatics assessment contexts cannot be assumed because considerable variations may exist between raters, targeted features, and examinee proficiency levels.
- Researchers should focus on the interpretive, rather than the descriptive, value in rating scale development, e.g.,
 - Category 2 (expressions that are incomprehensible, irrelevant, or too short) was found to be too narrowly defined in across several contingent factors.
 - Category 2 details various failed attempts; but does it represent a level of pragmatic ability distinctly different from the level represented by Category 1 (no attempt to respond)?

Thanks!



Keep in touch: sli12@gsu.edu



You can download the slides of this entire lecture series at:
https://scholarworks.gsu.edu/wcl_ilt/



Please cite this talk as

- Li, S. (2020. Nov. 14). Data collection methods & pragmatics assessment. Invited lecture series on L2 pragmatics (Lecture #5). Beijing Language and Culture University.